Neural Relation Extraction with Multi-lingual Attention

Yankai Lin[1], **Zhiyuan Liu**[1]*, **Maosong Sun**[1,2]
[1] Department of Computer Science and Technology,
State Key Lab on Intelligent Technology and Systems,
National Lab for Information Science and Technology, Tsinghua University, Beijing, China
[2] Jiangsu Collaborative Innovation Center for Language Competence, Jiangsu, China

Abstract

Relation extraction has been widely used for finding unknown relational facts from the plain text. Most existing methods focus on exploiting mono-lingual data for relation extraction, ignoring massive information from the texts in various languages. To address this issue, we introduce a multi-lingual neural relation extraction framework, which employs mono-lingual attention to utilize the information within mono-lingual texts and further proposes cross-lingual attention to consider the information consistency and complementarity among cross-lingual texts. Experimental results on real-world datasets show that our model can take advantage of multi-lingual texts and consistently achieve significant improvements on relation extraction as compared with baselines.

1 Introduction

People build many large-scale knowledge bases (KBs) to store structured knowledge about the real world, such as Wikidata[1] and DBpedia[2]. KBs are playing an important role in many AI and NLP applications such as information retrieval and question answering. The facts in KBs are typically organized in the form of triplets, e.g., (*New York*, `CityOf`, *United States*). Since existing KBs are far from complete and new facts are growing infinitely, meanwhile manual annotation of these knowledge is time-consuming and human-intensive, many works have been devoted to automated extraction of novel facts from various Web resources, where relation extraction (RE) from plain texts is one the most important knowledge sources.

Among various methods for relation extraction, distant supervision is the most promising approach (Mintz et al., 2009; Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012), which can automatically generate training instances via aligning KBs and texts to address the issue of lacking supervised data. As the development of deep learning, Zeng et al. (2015) introduce neural networks to extract relations with automatically learned features from training instances. To address the wrong labelling issue of distant supervision data, Lin et al. (2016) further employ sentence-level attention mechanism in neural relation extraction, and achieves the state-of-the-art performance.

However, most RE systems concentrate on extracting relational facts from mono-lingual data. In fact, people describe knowledge about the world using various languages. And people speaking different languages also share similar knowledge about the world due to the similarities of human experiences and human cognitive systems. For instance, though *New York* and *United States* are expressed as 纽约 and 美国 respectively in Chinese, both Americans and Chinese share the fact that "*New York* is a city of *USA*."

It is straightforward to build mono-lingual RE systems separately for each single language. But if so, it won't be able to take full advantage of diverse information hidden in the data of various languages. Multi-lingual data will benefit relation extraction for the following two reasons: **1. Consistency**. According to the distant supervision data in our experiments[3], we find that over half of Chinese and English sentences are longer than 20 words, in which only several words are related to the re-

* Corresponding author: Zhiyuan Liu (liuzy@tsinghua.edu.cn).

[1] http://www.wikidata.org/

[2] http://wiki.dbpedia.org/

[3] The data is generated by aligning Wikidata with Chinese Baidu Baike and English Wikipedia articles, which will be introduced in details in the section of experiments.

Relation	City in
English	1. **New York is a city** in the northeastern **United States**.
Chinese	1. **纽约**位于美国纽约州东南部大西洋沿岸，**是美国第一大城市**及第一大港. (**New York** is in the United States New York and on the Atlantic coast of the southeast Atlantic, **is the largest city** and largest port **in the United States**.) 2. **纽约是美国**人口最多**的城市**. (**New York is** the most populous **city in the United States**)

Table 1: An example of Chinese sentences and English sentence about the same relational fact (*New York*, `CityOf`, *United States*). Important parts are highlighted with bold face.

lational facts. Take Table 1 for example. The first Chinese sentence has over 20 words, in which only "纽约" (New York) and "是美国第一大城市" (is the biggest city in the United States) actually directly reflect the relational fact `CityOf`. It is thus non-trivial to locate and learn these relational patterns from complicated sentences for relation extraction. Fortunately, a relational fact is usually expressed with certain patterns in various languages, and the correspondence of these patterns among languages is substantially consistent. The pattern consistency among languages provides us augmented clues to enhance relational pattern learning for relation extraction.

2. Complementarity. From our experiment data, we also find that 42.2% relational facts in English data and 41.6% ones in Chinese data are unique. Moreover, for nearly half of relations, the number of sentences expressing relational facts of these relations varies a lot in different languages. It is thus straightforward that the texts in different languages can be complementary to each other, especially from those resource-rich languages to resource-poor languages, and improve the overall performance of relation extraction.

To take full consideration of these issues, we propose **M**ulti-lingual Attention-based **N**eural **R**elation **E**xtraction (MNRE). We first employ a convolutional neural network (CNN) to embed the relational patterns in sentences into real-valued vectors. Afterwards, to consider the complementarity of all informative sentences in various languages and capture the consistency of relational patterns, we apply mono-lingual attention to select the informative sentences within each language and propose cross-lingual attention to take advantages of pattern consistency and complementarity among languages. Finally, we classify relations according to the global vector aggregated from all sentence vectors weighted by mono-lingual attention and cross-lingual attention.

In experiments, we build training instances via distant supervision by aligning Wikidata with Chinese Baidu Baike and English Wikipedia articles, and evaluate the performance of relation extraction in both English and Chinese. The experimental results show that our framework achieves significant improvement for relation extraction as compared to all baseline methods including both mono-lingual and multi-lingual ones. It indicates that our framework can take full advantages of sentences in different languages and better capture sophisticated patterns expressing relations.

2 Related Work

Recent years KBs have been widely used on various AI and NLP applications. As an important approach to enrich KBs, relation extraction from plain text has attracted many research interests. Relation extraction typically classifies each entity pair into various relation types according to supporting sentences that the both entities appear, which needs human-labelled relation-specific training instances. Many works have been invested to relation extraction including kernel-based model (Zelenko et al., 2003), embedding-based model (Gormley et al., 2015), CNN-based models (Zeng et al., 2014; dos Santos et al., 2015), and RNN-based model (Socher et al., 2012).

Nevertheless, these RE systems are insufficient due to the lack of training data. To address this issue, Mintz et al. (2009) align plain text with Freebase to automatically generate training instances following the distant supervision assumption. To further alleviate the wrong labelling problem, Riedel et al. (2010) model distant supervision for relation extraction as a multi-instance single-label learning problem, and Hoffmann et al. (2011); Surdeanu et al. (2012) regard it as a multi-instance multi-label learning problem. Recently, Zeng et al. (2015) attempt to connect neural networks with distant supervision following the expressed-at-least-once assumption. Lin et al. (2016) further utilize sentence-level attention mechanism to consider all informative sentences jointly.

Most existing RE systems are absorbed in ex-

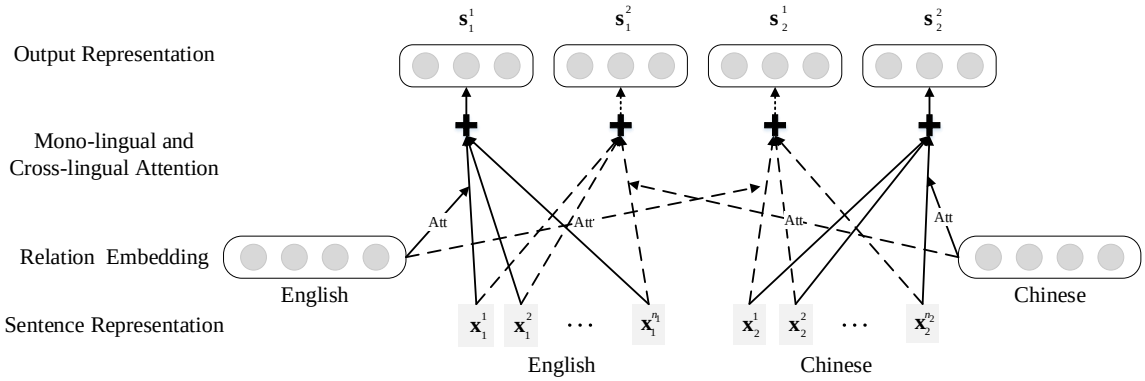

Figure 1: Overall architecture of our multi-lingual attention which contains two languages including English and Chinese. The solid lines indicates mono-lingual attention and the dashed lines indicates cross-lingual attention.

tracting relations from mono-lingual data, ignoring massive information lying in texts from multiple languages. In this area, Faruqui and Kumar (2015) present a language independent open domain relation extraction system, and Verga et al. (2015) further employ Universal Schema to combine OpenIE and link-prediction perspective for multi-lingual relation extraction. Both the works focus on multi-lingual transfer learning and learn a predictive model on a new language for existing KBs, by leveraging unified representation learning for cross-lingual entities. Different from these works, our framework aims to jointly model the texts in multiple languages to enhance relation extraction with distant supervision. To the best of our knowledge, this is the first effort to multi-lingual neural relation extraction.

The scope of multi-lingual analysis has been widely considered in many tasks besides relation extraction, such as sentiment analysis (Boiy and Moens, 2009), cross-lingual document summarization (Boudin et al., 2011), information retrieval in Web search (Dong et al., 2014) and so on.

3 Methodology

In this section, we describe our proposed MNRE framework in detail. The key motivation of MNRE is that, for each relational fact, the relation patterns in sentences of different languages should be substantially consistent, and MNRE can utilize the pattern consistency and complementarity among languages to achieve better results for relation extraction.

Formally, given two entities, their corresponding sentences in m different languages are defined as $T = \{S_1, S_2, \ldots, S_m\}$, where $S_j = \{x_j^1, x_j^2, \ldots, x_j^{n_j}\}$ corresponds to the sentence set in the jth language with n_j sentences. Our model measures a score $f(T, r)$ for each relation r, which is expected to be high when r is the valid one, otherwise low. The MNRE framework contains two main components:

1. Sentence Encoder. Given a sentence x and two target entities, we employ CNN to encode relation patterns in x into a distributed representation $\mathbf{x}$. The sentence encoder can also be implemented with GRU (Cho et al., 2014) or LSTM (Hochreiter and Schmidhuber, 1997). In experiments, we find CNN can achieve a better trade-off between computational efficiency and performance effectiveness. Thus, in this paper, we focus on CNN as the sentence encoder.

2. Multi-lingual Attention. With all sentences in various languages encoded into distributed vector representations, we apply mono-lingual and cross-lingual attentions to capture those informative sentences with accurate relation patterns. MNRE further aggregates these sentence vectors with weighted attentions into global representations for relation prediction.

We introduce the two components in detail as follows.

3.1 Sentence Encoder

The sentence encoder aims to transform a sentence x into its distributed representation $\mathbf{x}$ via CNN. First, it embeds the words in the input sentence into dense real-valued vectors. Next, it employs convolutional, max-pooling and non-linear transformation layers to construct the distributed representation of the sentence, i.e., $\mathbf{x}$.

3.1.1 Input Representation

Following (Zeng et al., 2014), we transform each input word into the concatenation of two kinds of representations: (1) a word embedding which captures syntactic and semantic meanings of the word, and (2) a position embedding which specifies the position information of this word with respect to two target entities. In this way, we can represent the input sentence as a vector sequence $\mathbf{w} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots\}$ with $\mathbf{w}_i \in \mathbb{R}^d$, where $d = d^a + d^b \times 2$. (d^a and d^b are the dimensions of word embeddings and position embeddings respectively)

3.1.2 Convolution, Max-pooling and Non-linear Layers

After encoding the input sentence, we use a convolutional layer to extract the local features, max-pooling, and non-linear layers to merge all local features into a global representation.

First, the convolutional layer extracts local features by sliding a window of length l over the sentence and perform a convolution within each sliding window. Formally, the output of convolutional layer for the ith sliding window is computed as:

$$\mathbf{p}_i = \mathbf{W}\mathbf{w}_{i-l+1:i} + \mathbf{b}, \tag{1}$$

where $\mathbf{w}_{i-l+1:i}$ indicates the concatenation of l word embeddings within the i-th window, $\mathbf{W} \in \mathbb{R}^{d^c \times (l \times d)}$ is the convolution matrix and $\mathbf{b} \in \mathbb{R}^{d^c}$ is the bias vector. (d^c is the dimension of output embeddings of the convolution layer)

After that, we combines all local features via a max-pooling operation and apply a hyperbolic tangent function to obtain a fixed-sized sentence vector for the input sentence. Formally, the ith element of the output vector $\mathbf{x} \in \mathbb{R}^{d^c}$ is calculated as:

$$[\mathbf{x}]_j = \tanh\big(\max_i(\mathbf{p}_{ij})\big). \tag{2}$$

The final vector $\mathbf{x}$ is expected to efficiently encode relation patterns about target entities from the input sentence.

Here, instead of max pooling operation, we can use piecewise max pooling operation adopted by PCNN (Zeng et al., 2015) which is a variation of CNN to better capture the relation patterns in the input sentence.

3.2 Multi-lingual Attention

To exploit the information of the sentences from all languages, our model adopts two kinds of attention mechanisms for multi-lingual relation extraction, including: (1) the mono-lingual attention which selects the informative sentences within one language and (2) the cross-lingual attention which measures the pattern consistency among languages.

3.2.1 Mono-lingual Attention

To address the wrong-labelling issue in distant supervision, we follow the idea of sentence-level attention (Lin et al., 2016) and set mono-lingual attention for MNRE. It is intuitive that each human language has its own characteristics. Hence we adopt different mono-lingual attentions to de-emphasize those noisy sentences within each language.

More specifically, for the j-th language and the sentence set S_j, we aim to aggregate all sentence vectors into a real-valued vector $\mathbf{S}_j$ for relation prediction. The mono-lingual vector $\mathbf{S}_j$ is computed as a weighted sum of those sentence vectors $\mathbf{x}_j^i$:

$$\mathbf{S}_j = \sum_i \alpha_j^i \mathbf{x}_j^i, \tag{3}$$

where α_j^i is the attention score of each sentence vector $\mathbf{x}_j^i$, defined as:

$$\alpha_j^i = \frac{\exp(e_j^i)}{\sum_k \exp(e_j^k)}, \tag{4}$$

where e_j^i is referred as a query-based function which scores how well the input sentence x_j^i reflects its labelled relation r. There are many ways to obtain e_j^i, and here we simply compute e_i as the inner product:

$$e_j^i = \mathbf{x}_j^i \cdot \mathbf{r}_j. \tag{5}$$

Here $\mathbf{r}_j$ is the query vector of the relation r with respect to the j-th language.

3.2.2 Cross-lingual Attention

Besides mono-lingual attention, we propose cross-lingual attention for neural relation extraction to better take advantages of multi-lingual data.

The key idea of cross-lingual attention is to emphasize those sentences which have strong consistency among different languages. On the basis of mono-lingual attention, cross-lingual attention is capable of further removing unlikely sentences and resulting in more concentrated and informative sentences, with the factor of consistent correspondence of relation patterns among different languages.

Cross-lingual attention works similar to mono-lingual attention. Suppose j indicates a language and k is a another language ($k \neq j$). Formally, the cross-lingual representation $\mathbf{S}_{jk}$ is defined as a weighted sum of those sentence vectors $\mathbf{x}_j^i$ in the jth language:

$$\mathbf{S}_{jk} = \sum_i \alpha_{jk}^i \mathbf{x}_j^i, \quad (6)$$

where α_{jk}^i is the cross-lingual attention score of each sentence vector $\mathbf{x}_j^i$ with respect to the kth language. The cross-lingual attention α_{jk}^i is defined as:

$$\alpha_{jk}^i = \frac{\exp(e_{jk}^i)}{\sum_k \exp(e_{jk}^k)}, \quad (7)$$

where e_{jk}^i is referred as a query-based function which scores the consistency between the input sentence x_j^i in the jth language and the relation patterns in the kth language for expressing the semantic meanings of the labelled relation r. Similar to the mono-lingual attention, we compute e_{jk}^i as follows:

$$e_{jk}^i = \mathbf{x}_j^i \cdot \mathbf{r}_k, \quad (8)$$

where $\mathbf{r}_k$ is the query vector of the relation r with respect to the kth language.

Note that, for convenience, we denote those mono-lingual attention vectors $\mathbf{S}_j$ as $\mathbf{S}_{jj}$ in the remainder of this paper.

3.3 Prediction

For each entity pair and its corresponding sentence set T in m languages, we can obtain $m \times m$ vectors $\{\mathbf{S}_{jk} | j, k \in \{1, \ldots, m\}\}$ from the neural networks with multi-lingual attention. Those vectors with $j = k$ are mono-lingual attention vectors, and those with $j \neq k$ are cross-lingual attention vectors.

We take all vectors $\{\mathbf{S}_{jk}\}$ together and define the overall score function $f(T, r)$ as follows:

$$f(T, r) = \sum_{j,k \in \{1,\ldots,m\}} \log p(r|\mathbf{S}_{jk}, \theta), \quad (9)$$

where $p(r|\mathbf{S}_{jk}, \theta)$ is the probability of predicting the relation r conditional on $\mathbf{S}_{jk}$, computed using a softmax layer as follows:

$$p(r|\mathbf{S}_{jk}, \theta) = \text{softmax}(\mathbf{M}\mathbf{S}_{jk} + \mathbf{d}), \quad (10)$$

where $\mathbf{d} \in \mathbb{R}^{n_r}$ is a bias vector, n_r is the number of relation types and $\mathbf{M} \in \mathbb{R}^{n_r \times R^c}$ is a global relation matrix initialized randomly.

To better consider the characteristics of each human language, we further introduce $\mathbf{R}_k$ as the specific relation matrix of the kth language. Here we simply define $\mathbf{R}_k$ as composed by $\mathbf{r}_k$ in Eq. (8). Hence, Eq. (10) can be extended to:

$$p(r|S_{jk}, \theta) = \text{softmax}[(\mathbf{R}_k + \mathbf{M})\mathbf{S}_{jk} + \mathbf{d}], \quad (11)$$

where $\mathbf{M}$ encodes global patterns for predicting relations and $\mathbf{R}_k$ encodes those language-specific characteristics.

Note that, in the training phase, the vectors $\{\mathbf{S}_{jk}\}$ are constructed using Eq. (3) and (6) using the labelled relation. In the testing phase, since the relation is not known in advance, we will construct different vectors $\{\mathbf{S}_{jk}\}$ for each possible relation r to compute $f(T, r)$ for relation prediction.

3.4 Optimization

Here we introduce the learning and optimization details of our MNRE framework. We define the objective function as follows:

$$J(\theta) = \sum_{i=1}^{s} f(T_i, r_i), \quad (12)$$

where s indicates the number of all entity pairs with each corresponding to a sentence set in different languages, and θ indicates all parameters of our framework.

To solve the optimization problem, we adopt mini-batch stochastic gradient descent (SGD) to minimize the objective function. For learning, we iterate by randomly selecting a mini-batch from the training set until converge.

4 Experiments

We first introduce the datasets and evaluation metrics used in the experiments. Next, we use a validation set to determine the best model parameters and choose the best model via early stopping. Afterwards, we show the effectiveness of our framework of considering pattern complementarity and consistency for multi-lingual relation extraction by quantitative and qualitative analysis. Finally, we compare the effect of two kinds of relation matrices in Eq. (11) used for prediction.

4.1 Datasets and Evaluation Metrics

We generate a new multi-lingual relation extraction dataset to evaluate our MNRE framework.

Without loss of generality, the experiments focus on relation extraction from two languages including English and Chinese. In this dataset, the Chinese instances are generated by aligning Chinese Baidu Baike with Wikidata, and the English instances are generated by aligning English Wikipedia articles with Wikidata. The relational facts of Wikidata in this dataset are divided into three parts for training, validation and testing respectively. There are 176 relations including a special relation NA indicating there is no relation between entities. And we set both validation and testing sets for Chinese and English parts contain the same facts. We list the statistics about the dataset in Table 2.

Dataset		#Rel	#Sent	#Fact
English	Train	176	1,022,239	47,638
	Valid		80,191	2,192
	Test		162,018	4,326
Chinese	Train	176	940,595	42,536
	Valid		82,699	2,192
	Test		167,224	4,326

Table 2: Statistics of the dataset.

We follow previous works (Mintz et al., 2009) and investigate the performance of RE systems using the held-out evaluation, by comparing the relational facts discovered by RE systems from the testing set with those facts in KB. The evaluation method assumes that if a RE system accurately finds more relational facts in KBs from the testing set, it will achieve better performance for relation extraction. The held-out evaluation provides an approximate measure of RE performance without time-consuming human evaluation. In experiments, we report the precision/recall curves as the evaluation metric.

4.2 Experimental Settings

We tune the parameters of our MNRE framework by grid searching using validation set. For training, we set the iteration number over all the training data as 15. The best models were selected by early stopping using the evaluation results on the validation set. In Table 3 we show the best setting of all parameters used in our experiments.

4.3 Effectiveness of Consistency

To demonstrate the effectiveness of considering pattern consistency among languages, we empirically compare different methods through held-out evaluation. We select CNN proposed in (Zeng et al., 2014) as our sentence encoder and implement it by ourselves which achieves comparable results as the authors reported on their experimental dataset NYT10[4]. And we compare the performance of our framework with the [P]CNN model trained with only English data ([P]CNN-En), only Chinese data ([P]CNN-Zh), a joint model ([P]CNN+joint) which predicts using [P]CNN-En and [P]CNN-Zh jointly, and another joint model with shared embeddings ([P]CNN+share) which trains [P]CNN-En and [P]CNN-Zh with common relation embedding matrices.

Hyper-parameter	value
Window size w	3
Sentence embedding size d^c	230
Word dimension d^a	50
Position dimension d^b	5
Batch size B	160
Learning rate λ	0.001
Dropout probability p	0.5

Table 3: Parameter settings.

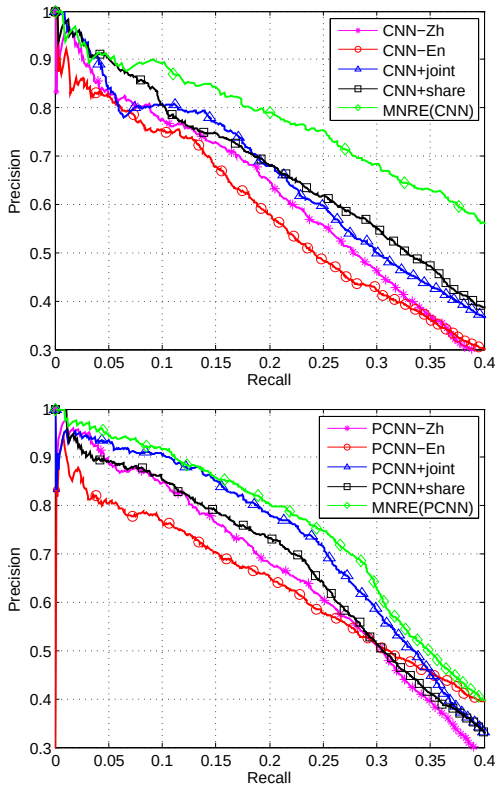

Figure 2: Top: Aggregated precision/recall curves of CNN-En, CNN-Zh, CNN+joint, CNN+share, and MNRE(CNN). Bottom: Aggregated precision/recall curves of PCNN-En, PCNN-Zh, PCNN+joint, PCNN+share, and MNRE(PCNN)

[4] http://iesl.cs.umass.edu/riedel/ecml/

CNN+Zh	CNN+En	MNRE	Sentence
—	Medium	Low	1. **Barzun** is a commune in the Pyrénées-Atlantiques department in the Nouvelle-Aquitaine region of south-western **France**.
—	Medium	High	2. **Barzun** was born in Créteil , France
Medium	—	Low	3. 作为从**法国**移民到美国来的顶尖知识分子，**巴尔赞**与莱昂内尔·特里林、德怀特·麦克唐纳等人一道，在冷战时期积极参与美国的公共知识生活 …(As a top intellectual immigrating from **France** to the United States, **Barzun**, together with Lionel Trilling and Dwight Macdonald, actively participated in public knowledge life in the United States during the cold war …)
Medium	—	High	4. **巴尔赞**于 1907 年出生于法国一个知识分子家庭，1920 年赴美。(**Barzun** was born in a **French** intellectual family in 1907 and went to America in 1920.)

Table 4: An example of our multi-lingual attention. Low, medium and high indicate the attention weights.

From Fig. 2, we have the following observations:

(1) Both [P]CNN+joint and [P]CNN+share achieve better performances as compared to [P]CNN-En and [P]CNN-Zh. It indicates that utilizing Chinese and English sentences jointly is beneficial to extracting novel relational facts. The reason is that those relational facts that are discovered from multiple languages are more reliable to be true.

(2) CNN+share only has similar performance as compared to CNN+joint, even through a bit worse when recall ranges from 0.1 to 0.2. Besides, PCNN+share performs worse than PCNN+joint nearly over the entire range of recall. It demonstrates that a simple combination of multiple languages by sharing relation embedding matrices cannot further capture more implicit correlations among various languages.

(3) Our MNRE model achieves the highest precision over the entire range of recall as compared to other methods including [P]CNN+joint and [P]CNN+share models. By grid searching of parameters for these baseline models, we can observe that both [P]CNN+joint and [P]CNN+share cannot achieve competitive results compared to MNRE even when increasing the size of the output layer. This indicates that no more useful information can be captured by simply increasing model size. On the contrary, our proposed MNRE model can successfully improve multi-lingual relation extraction by considering pattern consistency among languages.

We further give an example of cross-lingual attention in Table 4. It shows four sentences having the highest and lowest Chinese-to-English and English-to-Chinese attention weights respectively with respect to the relation `PlaceOfBirth` in MNRE. We highlight the entity pairs in bold face. For comparison, we also show their attention weights from CNN+Zh and CNN+En. From the table we find that, although all of the four sentences actually express the fact that Barzun was born in France, the first and third sentences contain much more noisy information that may confuse RE systems. By considering pattern consistency between sentences in two languages with cross-lingual attention, MNRE can identify the second and fourth sentences that unambiguously express the relation `PlaceOfBirth` with higher attention as compared to CNN+Zh and CNN+En.

4.4 Effectiveness of Complementarity

To demonstrate the effectiveness of considering pattern complementarity among languages, we empirically compare the following methods through held-out evaluation: MNRE for English (MNRE-En) and MNRE for Chinese (MNRE-Zh) which only use the mono-lingual vectors to predict relations, and [P]CNN-En and [P]CNN-Zh models.

Fig. 3 shows the aggregated precision/recall curves of the four models for both CNN and PCNN. From the figure, we find that:

(1) MNRE-En and MNRE-Zh outperform [P]CNN-En and [P]CNN-Zh almost in entire range of recall. It indicates that by jointly training with multi-lingual attention, both Chinese and English relation extractors are beneficial from those sentences from the other language.

(2) Although [P]CNN-En underperforms as compared to [P]CNN-Zh, MNRE-En is comparable to MNRE-Zh by jointly training through multi-lingual attention. It demonstrates that both Chinese and English relation extractors can take full advantages of texts in both languages via our propose multi-lingual attention scheme.

Table 5 shows the detailed results (in precision@1) of some specific relations of which the training instances are un-balanced on English and

Relation	#Sent-En	#Sent-Zh	CNN-En	CNN-Zh	MNRE-En	MNRE-Zh
Contains	993	6984	17.95	69.87	73.72	75.00
HeadquartersLocation	1949	210	43.04	0.00	41.77	50.63
Father	1833	983	64.71	77.12	86.27	83.01
CountryOfCitizenship	25322	15805	95.22	93.23	98.41	98.21

Table 5: Detailed results (precision@1) of some specific relations. #Sent-En and #Sent-Zh indicate the numbers of English/Chinese sentences which are labelled with the relations.

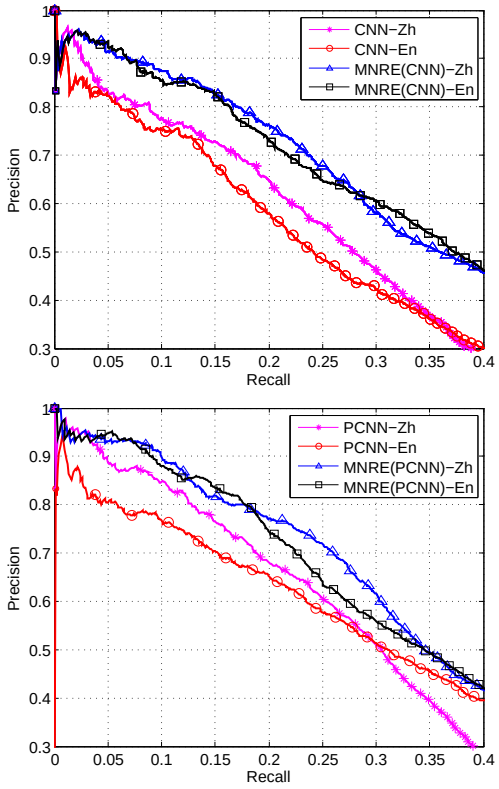

Figure 3: Top: Aggregate precision/recall curves of CNN-En, CNN-Zh, MNRE(CNN)-En and MNRE(CNN)-Zh. Bottom: Aggregate precision/recall curves of PCNN-En, PCNN-Zh, MNRE(PCNN)-En and MNRE(PCNN)-Zh.

Chinese sides. From the table, we can see that:

(1) For the relation Contains of which the number of English training instances is only $1/7$ of Chinese ones, CNN-En gets much worse performance as compared to CNN-Zh due to the lack of training data. Nevertheless, by jointly training through multi-lingual attention, MNRE(CNN)-En is comparable to and slightly better than MNRE(CNN)-Zh.

(2) For the relation HeadquartersLocation of which the number of Chinese training instances is only $1/9$ of English ones, CNN-Zh even cannot predict any correct results. The reason is perhaps that, CNN-Zh of the relation is not sufficiently trained because there are only 210 Chinese training instances for this relation. Similarly, by jointly training through multi-lingual attention, MNRE(CNN)-En and MNRE(CNN)-Zh both achieve promising results.

(3) For the relations Father and CountryOfCitizenship of which the sentence number in English and Chinese are not so un-balanced, our MNRE can still improve the performance of relation extraction on both English and Chinese sides.

4.5 Comparison of Relation Matrix

For relation prediction, we use two kinds of relation matrices including: **M** that considers the global consistency of relations, and **R** that considers the specific characteristics of relations for each language. To measure the effect of the two relation matrices, we compare the performance of MNRE using the both matrices with those only using **M** (MNRE-M) and only using **R** (MNRE-R).

Fig. 4 shows the precision-recall curves for each method. From the figure, we observe that:t

(1) The performance of MNRE-M is much worse than both MNRE-R and MNRE. It indicates that we cannot just use global relation matrix for relation prediction. The reason is that each language has its own specific characteristics to express relation patterns, which cannot be well integrated into a single relation matrix.

(2) MNRE(CNN)-R has similar performance as compared to MNRE(CNN) when the recall is low. However, it has a sharp decline when the recall reaches 0.25. It suggests there also exists global consistency of relation patterns among languages which cannot be neglected. Hence, we should combine both **M** and **R** together for multi-lingual relation extraction, as proposed in our MNRE framework.

5 Conclusion

In this paper, we introduce a neural relation extraction framework with multi-lingual attention to take pattern consistency and complementarity among

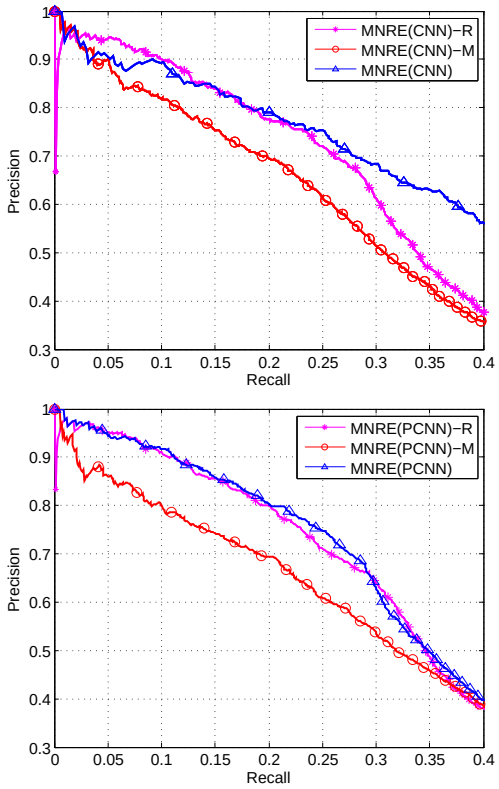

Figure 4: Top: Aggregated precion/recall curves of MNRE(CNN)-M, MNRE(CNN)-R and MNRE. Bottom: Aggregated precion/recall curves of MNRE(PCNN)-M, MNRE(PCNN)-R and MNRE(PCNN).

multiple languages into consideration. We evaluate our framework on multi-lingual relation extraction task, and the results show that our framework can effectively model relation patterns among languages and achieve state-of-the-art results.

We will explore the following directions as future work:

(1) In this paper, we only consider sentence-level multi-lingual attention for relation extraction. In fact, we find that the word alignment information may be also helpful for capturing relation patterns. Hence, the word-level multi-lingual attention, which may discover implicit alignments between words in multiple languages, will further improve multi-lingual relation extraction. We will explore the effectiveness of word-level multi-lingual attention for relation extraction as our future work.

(2) MNRE can be flexibly implemented in the scenario of multiple languages, and this paper focuses on two languages of English and Chinese. In future, we will extend MNRE to more languages and explore its significance.

Acknowledgments

This work is supported by the 973 Program (No. 2014CB340501), the National Natural Science Foundation of China (NSFC No. 61572273, 61532010), and the Key Technologies Research and Development Program of China (No. 2014BAK04B03). This work is also funded by the Natural Science Foundation of China (NSFC) and the German Research Foundation(DFG) in Project Crossmodal Learning, NSFC 61621136008 / DFC TRR-169.

References

Erik Boiy and Marie-Francine Moens. 2009. A machine learning approach to sentiment analysis in multilingual web texts. *Information retrieval* 12(5):526–558.

Florian Boudin, Stéphane Huet, and Juan-Manuel Torres-Moreno. 2011. A graph-based approach to cross-language multi-document summarization. *Polibits* (43):113–118.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* .

Meiping Dong, Yong Cheng, Yang Liu, Jia Xu, Maosong Sun, Tatsuya Izuha, and Jie Hao. 2014. Query lattice for translation retrieval. In *Proceedings of COLING*. pages 2031–2041.

Cıcero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. In *Proceedings of ACL*. volume 1, pages 626–634.

Manaal Faruqui and Shankar Kumar. 2015. Multilingual open relation extraction using cross-lingual projection. *arXiv preprint arXiv:1503.06450* .

Matthew R. Gormley, Mo Yu, and Mark Dredze. 2015. Improved relation extraction with feature-rich compositional embedding models. In *Proceedings of EMNLP*. pages 1774–1784.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* pages 1735–1780.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of ACL-HLT*. pages 541–550.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of ACL*. volume 1, pages 2124–2133.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL-IJCNLP*. pages 1003–1011.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of ECML-PKDD*. pages 148–163.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of EMNLP-CoNLL*. pages 1201–1211.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *JMLR* 15(1):1929–1958.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of EMNLP*. pages 455–465.

Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. 2015. Multilingual relation extraction using compositional universal schema. *arXiv preprint arXiv:1511.06396* .

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *JMLR* 3(Feb):1083–1106.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of EMNLP*.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING*. pages 2335–2344.